UAMCLyR at RepLab 2014: Author Profiling Task*

Notebook for RepLab at CLEF 2014

E. Villatoro-Tello, G. Ramírez-de-la-Rosa, C. Sánchez-Sánchez,
H. Jiménez-Salazar, W. A. Luna-Ramírez, and C. Rodríguez-Lucatero

Departamento de Tecnologías de la Información,
Universidad Autónoma Metropolitana, Unidad Cuajimalpa,
Ave. Vasco de Quiroga Num. 4871 Col Santa Fe, México D.F.
{evillatoro,gramirez,csanchez,hjimenez,wluna,crodriguez}@correo.cua.uam.mx

Abstract. This paper describes the participation of the Language and Reasoning Group of UAM at RepLab 2014 Author Profiling evaluation lab. This task involves *author categorization* and *author ranking* subtasks. Our method for *author categorization* uses a supervised approach based on the idea that we can use the information on Twitter's user profile, then by means of employing an attribute selection techniques we can extract attributes that are the most representative from each user's activity domain. For the *author ranking* subtask we use a two step chained method that uses *stylistics* attributes (e.g. lexical richness, language complexity) and *behavioral* attributes (e.g. posts' frequency, directed tweets) extracted from the users' profile and the posts. We use these attributes in conjunction with a Markov Random Fields for improving an initial ranking given by the confidence of Support Vector Machine classification algorithm. Obtained results are encouraging and motivate us to keep working on the same ideas.

Keywords: Author Profiling, Supervised Text Classification, Probabilistic Ranking Method, Markov Random Field, Communication Behavior

1 Introduction

From its inception in 2006, Twitter has become one of the most important platform for microblog posts. Recent statistics reveal that there are more that 250 million users that write more than 500 million posts every day[1], talking about a great diversity of topics. As a consequence, several entities such as companies, celebrities, politicians, etc., are very interested in using this type of platform for increasing or even improving their presence among Twitter users, aiming at obtaining good reputation values.

As an important effort for providing effective solutions to the above problem, RepLab[2] proposes a competitive evaluation exercise for Online Reputation Management (ORM) systems. For this year RepLab campaign (RepLab 2014 [1]), the *Author Profiling* task was one of the main evaluated tasks.

* This work was partially supported by CONACyT México Project Grant CB-2010/153315, and SEP-PROMEP UAM-PTC-380/48510349.

[1] *https://about.twitter.com/company*

[2] *http://www.limosine-project.eu/events/replab2013*

The Author Profiling task consisted on two subtasks, namely the *Author Categorization* and *Author Ranking* subtasks. On the one hand, the *author categorization* subtask consists in detecting author's activity domain, *e.g.*, discovering if certain author is a journalist, an activist, etc. On the other hand, the *author ranking* subtask consists in discovering those users that could represent an opinion leader among a community, *i.e.*, finding those users who are the most influential (opinion makers) within a community of users.

In recent years an increasing number of methods and systems are been developed to tackle the author profiling task. Most of these methods deal with profiling long texts, for instance, posts in a blog [8,11], conversations in a chatroom [3], etc. In these scenarios, systems usually count with enormous collections of manually labeled examples, where a set of profiles categories are usually known a priori. However, the task in the RepLab is a bit different from these previous systems in that the possible set of profile categories is unknown, thus the lack of examples in some categories adds complexity to this year's challenge.

The author ranking subtask is as new as the social media has been. One of the challenges here is to identify an author or authors that are influential to a particular community. For Twitter, some methods [7] use a variation of the PageRank algorithm, taking advantage of the following-followers schema on Twitter. Some others methods [5], set the ranking of an author according to the number of tweets that are important. However, in a very dynamic environment, some of these methods may have some difficulties updating the ranks for every user or even more every tweet.

Our proposed approach for facing the problem of *author categorization* is based on the idea that information in the user's profile description is good enough to find the author's activity domain. To accomplish this task our method uses different attribute selection techniques to find the most representative characteristics, namely words, for each activity domain. A variation for this approach employs a term expansion technique that aims at improving the descriptive terms in each user's profile.

Additionally, our proposed approach for *author ranking* problem is based on the idea that the rank (leadership) of an author can be detected by considering its writing style, and its behavior within the Twitter's community. To accomplish this task, our method uses different stylistics attributes (lexical richness, language complexity, etc), as well as some behavioral features (posts' frequency, directed tweets, etc).

Accordingly, this paper describes the participation of the Language and Reasoning research group from UAM-C to the CLEF 2014 RepLab author profiling task (*i.e.*, *author categorization* and *author ranking* subtasks) [4,1]. The main objectives of our experiments were:

1. *Determine if it is possible to categorize author based on solely the information that every Twitter's user share in the profile information.*
2. *Determine if it is possible to rank an user based on stylistics and behavioral attributes extracted from their posts.*

The rest of this paper is organized as follows. The next section describes all the steps considered in developing and performing all of our experiments for the *author categorization* subtask. Then, Section 3 describe the proposed approaches for solving

the problem of *author ranking*. Finally, Section 4 presents the conclusions derived from this work and outlines future work directions.

2 Author Categorization

The Author Categorization subtask consisted in classifying Twitter profiles by type of author (*i.e,* journalist, professional, authority, activist, investor, company or celebrity). Since we were given the main categories on the training set (*i.e., twitter profiles*), we faced the problem as a supervised approach, particularly as a Text Classification (TC) task. As a first step, all documents contained in the training and test sets were preprocessed: deleting stop words, URLs and using a Porter stemmer algorithm [9]. Next, we computed the Document Frequency (*DF*) score for each term contained in the training collection. Then, terms were sorted according to their frequency values (*i.e., DF* values). Finally, we applied a term selection technique, which consists of selecting sets from 10% to 100% of terms according to their *DF* value.

We selected as main classifier method the Support Vector Machines (SVM) approach as implemented in Weka. Results are reported in terms of *Precision*, *Recall* and *F-score*, while as a validation strategy we employed a *10 cross-fold-validation* technique.

Table 1. Obtained results on the training set when terms are selected considering their *DF* frequency. Notice that the best results was obtained when using the 80% of the terms.

Selected terms (%)	Precision	Recall	F-score
10%	0.45	0.48	0.45
20%	0.45	0.48	0.46
30%	0.46	0.49	0.47
40%	0.46	0.50	0.47
50%	0.46	0.50	0.47
60%	0.47	0.50	0.47
70%	0.47	0.50	0.47
80%	0.48	0.51	**0.48**
90%	0.47	0.50	0.47

Table 1 shows obtained results on the training set. As can be noticed, the best result was obtained when only the 80% of *DF* terms are employed to represent the Twitter's profiles.

Results obtained in our first experiment (Table 1) indicate, to some extent, that contained words in Twitter profiles are able to provide some information that allows traditional text classification techniques to get an acceptable performance. Consequently, for our second experiment (Table 2) we considered expanding profiles information. In order to perform a profile expansion, we search for synonyms of every word contained in user's profiles. To accomplish this we use WordNet as main resource. After expan-

sion is performed, we compute the DF score of all terms and follow a similar process to the one described above.

Table 2. Obtained results on the training set when users' profiles are expanded by means of WordNet. Notice that the best results are obtained when documents are represented using 90% of terms.

Selected terms (%)	Precision	Recall	F-score
10%	0.40	0.42	0.41
20%	0.40	0.42	0.41
30%	0.41	0.43	0.42
40%	0.41	0.44	0.42
50%	0.42	0.44	0.43
60%	0.43	0.46	0.44
70%	0.43	0.46	0.44
80%	0.43	0.46	0.44
90%	0.44	0.47	**0.45**

Table 2 shows obtained results when users' profiles are expanded by means of adding synonyms extracted from WordNet. Similarly to our first experiment, we used a SVM classification method. Notice, that the classifier's performance decreased in comparison to results obtained in Table 1. Observed behavior forces us to think that expanding all terms is being harmful since it adds more noisy elements.

Finally, as our third and fourth set of experiments we applied as a term selection strategy one of the methods that has been proved effective in thematic TC, namely the transition point (pt_T) [10]. The tp_T represents a frequency value that accurately divides the vocabulary in two subsets, those of low frequency and high frequency [13]. Some empirical results [6] indicate that by means of preserving those terms surrounding the tp_T it is possible to solve (to some extent) some non-thematic TC tasks.

Consequently, for our third and fourth experiments (second and third row from Table 3) we applied the tp_T instead of the DF as term selection strategy. Table 3 shows obtained results on the training set when this strategy is employed.

Table 3. Obtained results on the training set when the transition point technique is employed as term selection strategy. For both experiments only the 40% of terms surrounding the transition point were preserved.

Terms Expansion	Precision	Recall	F-score
NO	0.39	0.43	0.40
YES	0.39	0.42	0.39

The first column in Table 3 indicates whether we applied or not a terms expansion technique on users' profiles. As can be observed, better results are obtained when no expansion is performed.

2.1 Submitted runs

Based on the results obtained on the training set, we define the following as our official experiments.

UAM-CALYR-AC-1 : This experiment uses as a term selection strategy the *DF* score. According with experimental results (Table 1) we represented users' profiles by means of the 80% terms whit higher *DF* scores.

UAM-CALYR-AC-2 : Similar to previous experiment, this configuration represents users' profiles with the 80% of terms. The main difference is that this experiments applies a users' profiles expansion using WordNet.

UAM-CALYR-AC-3 : The transition point strategy is employed as term selection strategy. No users' profiles expansion is performed, and only the 40% of terms surrounding the tp_T are preserved.

UAM-CALYR-AC-4 : The transition point strategy is employed as term selection strategy. Users' profiles expansion is performed by means of WordNet, and only the 40% of terms surrounding the tp_T are preserved.

Table 4 shows obtained official results during RepLab 2014 campaign. It is worth mentioning that UAM-CALYR-AC-3 and UAM-CALYR-AC-4 were submitted as unofficial runs. Last three rows of Table 4 depict the obtained performance of the official baselines proposed by RepLab organizers (*i.e, All in one class, and baseline SVM*). The last row (*All systems average*) represents the Macro-average performance obtained by all participant groups. Results are reported in terms of *Reliability*, *Sensitivity* and *F* measure [2].

Table 4. Obtained results on the test set for *author categorization* subtask.

Run ID	Reliability	Sensitivity	F
UAM-CALYR-AC-1	0.36	0.37	0.36
UAM-CALYR-AC-2	0.35	0.40	0.36
UAM-CALYR-AC-3	0.28	1.00	0.44
UAM-CALYR-AC-4	0.28	1.00	0.44
All in one class	0.28	1.00	0.44
baseline-SVM	0.19	0.35	0.25
All systems average	0.29	0.46	0.33

As can be observed in Table 4 our proposed experiments are able to achieve a competitive performance. Notice that in terms of the *F* measure, our methods that use the tp_T as a term selection strategy are able to reach a 0.44 score. Generally speaking, proposed approaches are able to obtain good levels of *reliability* compared to proposed

baselines and systems average, indicating to some extent that proposed approaches allow better *precision* values without decreasing *sensitivity* values.

3 Author Ranking

We approached the *Author Ranking* problem as two-step chained method. The first phase is a supervised approach, and the later a unsupervised approach that uses a Markov Random Field. The general schema of our proposed method is shown in Figure 1 and a detail description of each steps is as following.

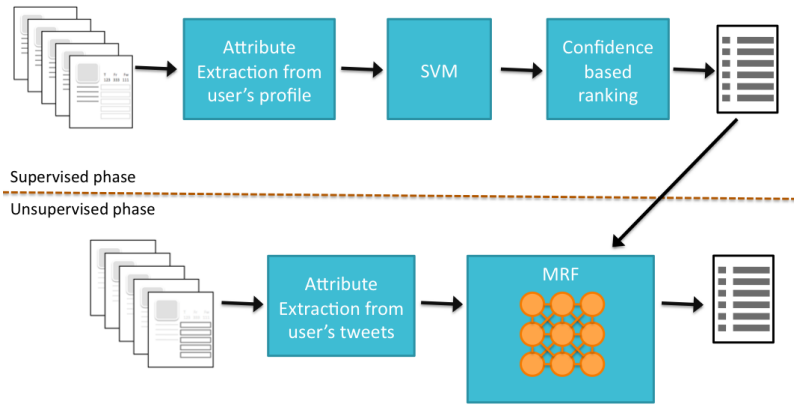

Fig. 1. General Schema of our proposed method for Author Ranking

The *supervised phase* has as input two sets of attributes that aim at capturing stylistics and behavioral characteristics from each user. Each set is composed as follows:

- *Self-description (SD)*. Words in the user's profile plus tags we use for mentions to users (mentions), number of hashtags, URLs, and Twitter ID's user appearing in the profile (user-info).
- *Statistics of use (SU)*. Number of total tweets, number of followers, number of following, average of tweets per followers, ratio of following by followers, ratio of followers by following, and author's category activity.

Our intuition is that opinion maker users tend to write profile descriptions in a more professional fashion than non opinion makers. Also, they tend to link to external URL that usually contains the user name on their Twitter ID. We also believe that influential users follow fewer people than people following them.

Once we extracted the set of attributes, we tackled the ranking problem as a binary classification task (with classes *opinion maker* and *non opinion maker*). We use a SVM to learn the model, and the final ranked list is generated according to the confidence of the SVM classification algorithm.

For our experiments, we wanted to find if the information of only the supervised phase is enough for providing a proper author ranking. In this direction we propose two experiments: *i)* using the complete set of extracted attributes (*self-description* and *statistics of use*), and *ii*) using only *statistics of use*. Table 5 shows the results obtained over the training set. According to these results, self descriptions add some information that is relevant to correctly classify a user into opinion maker and non-opinion maker categories so the rank can be assigned correctly.

Table 5. Obtained results on the training set: Supervised phase.

Experiment ID	P@10	P@20	R-Prec	MAP
SVM: SD+SU	0.60	0.5750	0.6574	**0.5717**
SVM: SU	0.60	0.5750	0.5703	0.5159

For the unsupervised phase, the ranking is obtained based on a Markov Random Field (MRF) model that separates the opinion maker users from non-opinion makers, generating a new list by positioning the relevant users (opinion makers) first, and the others after.

MRF are a type of undirected probabilistic graphical models that aim at modeling dependencies among variables of the problem in turn. MRF modeling has appealing features for problems that involve the optimization of a configuration of variables that have interdependencies among them. Accordingly, MRFs allow the incorporation of contextual information in a principled way. MRFs rely on a strict probabilistic modeling, yet they allow the incorporation of prior knowledge by means of potential functions. For those reasons, in this work we employed an MRF model for refining the initial ranking of a set of users previously ordered by a supervised approach (Figure 1).

For our performed experiments, we adopted a MRF configured as described in [12]. Accordingly, in our considered MRF each node corresponds to a user in the list. Each user is represented as a random variable with 2 possible values: *opinion maker* and *non-opinion maker*. Similarly to [12] we consider a fully connected graph, such that each node (user) is connected to all other nodes in the field; that is, we defined a neighbourhood scheme in which each variable is adjacent to all the others.

Contrary to [12], we only considered the information provided by the interaction potential[3], which asses how much support provide the neighbouring same-valued variables to some particular node f_i so it keep its current value, and also how much support give oppose-value variables to f_i so it changes to the contrary value.

Additionally, for estimating similarities among users we did not use textual features, instead we propose a set of *stylistic* and *behavioral* features that are extracted from users' posts which are described below:

[3] Originally, the work showed in [12] considers two potentials: *i)* interaction potential which accounts for information of the association between surrounding nodes, and *ii)* observation potential that accounts for information that is associated to a single node.

– *Style-Behavior (SB).* In order to determine the set of *SB* features we compute for each tweet from each user the following features: number of URLs, hashtags, user mentions, number of employed words, average size of words, user name length, vocabulary richness, hapax's number, average number of retweets, favorites, number of characters per tweet, number of employed special symbols (not words), size of user mentions and hashtags (in characters) and the average posting frequency time with their respective standard deviation. At the end, all calculations are averaged so that each user is represented by a vector of 16 elements.

Our intuitive idea for proposing *SB* features is that *opinion maker* users would have similar writing styles as well as similar posting behaviors. Hence, the optimal configuration of the MRF would be able to gather both opinion-maker users and non-opinion maker users.

Finally, the initial configuration of the MRF is obtained by considering the output provided by the supervised phase (Figure 1). That is, the subset of users that were classified as *opinion maker* by the supervised phase are initialized as true *opinion makers* within the MRF, and all other users as *non-opinion makers*. Then, the MRF configuration of minimum energy (MAP) is obtained via stochastic simulation using the ICM algorithm. At the end of this optimization process, a new re-ordered (improved) list is produced.

Our second set of experiments are directed to evaluate the unsupervised phase, thus we can evaluate the contribution of proposed MRF at improving the initial author ranking (*i.e.,* improving the output of the supervised phase). To accomplish this goal, we design three more experiments: *i)* using the ranked list from experiment *SVM: SD+SU* as input for the MRF, and, for estimating similarities within the MRF users are represented by means of features *SD+SU+SB*, *ii)* using the ranked list from experiment *SVM: SU* as input for the MRF, and for similarities estimation users are represented by means of *SU+SB* features, *iii)* using the ranked list from experiment *SVM: SD+SU* as input for the MRF, and for similarities estimation users are represented by just *SB* features.

Table 6. Obtained results on the training set: Unsupervised phase.

Experiment ID	**P@10**	**P@20**	**R-Prec**	**MAP**
SVM: SD+SU & MRF: SD+SU+SB	0.60	0.65	0.6616	0.5965
SVM: SU & MRF: SU+SB	0.45	0.55	0.5833	0.5290
SVM: SD+SU & MRF: SB	0.60	0.65	0.6689	**0.6116**

As can be observed in Table 6 the MRF ranking proposal improves the initial ranking generated by the supervised phase (See Table 5). It is worth to remark that the best result was obtained when just the style and behavior (*SB*) features, are employed to estimate similarities within the MRF. These results indicate, to some extent, that opinion makers do have similar writing styles as well as similar behavior patterns when using Twitter. In addition, obtained results also indicate that features extracted from profile

information, *i.e.,* SD or SU features, are insufficient and even noisy for the MRF configuration, however the use of these attributes on a supervised fashion are a cheap and efficient option to assign an initial ranking for authors.

3.1 Submitted runs

Based on the results obtained on the training set, we define the following as our official experiments.

UAM-CALYR-AR-1 : Using the supervised phase with SVM learning method and attributes: self-description and statistics of use. That is, the same configuration of experiment *SVM:SD+SU*, see Table 5.

UAM-CALYR-AR-2 : Using the complete schema showed in Figure 1. SVM as learning method with SD and SU attributes, generated ranking of these serves as input for the MRF which employs SD+SU+SB for representing users when estimating similarities within the MRF. This experiment uses the same configuration as experiment *SVM: SD+SU & MRF: SD+SU+SB* from Table 6.

UAM-CALYR-AR-3 : Using the supervised phase only with SVM learning method and statistics of use as attributes. That is the same configuration of experiment *SVM:SU*, see Table 5.

UAM-CALYR-AR-4 : Using the complete schema showed in Figure 1. SVM as learning method with SU attributes, generated ranking of these serves as input for the MRF which employs SU+SB for representing users when estimating similarities within the MRF. This experiment uses the same configuration as experiment *SVM:SU & MRF:SU+SB* from Table 6.

UAM-CALYR-AR-4 : Using the complete schema showed in Figure 1. SVM as learning method with SD and SU attributes, generated ranking of these serves as input for the MRF which employs just SB features for representing users when estimating similarities within the MRF. This experiment uses the same configuration as experiment *SVM: SD+SU & MRF: SB* from Table 6.

Table 7. Obtained results on the test set for *author ranking* subtask considering the three domains: *Automotive, Banking* and *Miscellaneous*.

Run ID	**P@10**	**P@20**	**R-Prec**	**MAP**
UAM-CALYR-AR-1	0.70	0.63	0.5677	0.5464
UAM-CALYR-AR-2	0.70	0.63	0.5669	0.5461
UAM-CALYR-AR-3	0.70	0.70	0.4977	0.5139
UAM-CALYR-AR-4	0.70	0.70	0.4981	0.5137
UAM-CALYR-AR-5	0.70	0.63	0.5515	**0.5677**
Followers	0.73	0.66	0.5303	0.5530
All systems average	0.70	0.68	0.5224	0.5239

Table 7 shows official results obtained for our submitted experiments. The second-last row is the official baseline proposed by RepLab organizers (*i.e. Followers*). The

last row (*All systems average*) represents the Macro-average performance obtained by all participant groups.

It is worth mentioning that showed results in Table 7, represent the average MAP performance among the three domains *i.e., Automotive, Banking* and *Miscellaneous*, released as test set. As expected, our best result was achieved when the MRF uses just the *SB* features for estimating similarities among users' profiles. Hence, if we consider the three test domains, the MAP performance obtained by our experiment UAM-CALYR-AR-5 represents the second best system during RepLab 2014 campaign, particularly for the author ranking subtask.

Table 8. Obtained results on the test set for *author ranking* subtask considering just two domains: *Automotive* and *Banking*.

Run ID	**P@10**	**P@20**	**R-Prec**	**MAP**
UAM-CALYR-AR-1	0.6	0.5	0.5045	0.4361
UAM-CALYR-AR-2	0.6	0.5	0.5032	0.4355
UAM-CALYR-AR-3	0.7	0.6	0.3936	0.3813
UAM-CALYR-AR-4	0.7	0.6	0.3942	0.3809
UAM-CALYR-AR-5	0.6	0.5	0.4802	**0.4653**
Followers	0.6	0.5	0.3895	0.3739
All systems average	0.67	0.64	0.4729	0.4586

Table 8 shows official results when two domains are considered (*Automotive* and *Banking*). As it is possible to observe, our best experiment in terms of the MAP measure, was obtained by the UAM-CALYR-AR-5 experiment, *i.e.,* when the MRF uses just the *SB* features for estimating similarities among users' profiles. Under this type of evaluation *i.e.,* considering only two domains for evaluations, our best experiment (UAM-CALYR-AR-5) gets ranked at seventh place among participant systems.

4 Conclusion and Future work

In this paper, we have described the experiments performed by the Language and Reasoning group from UAM-C in the context of the RepLab 2014 evaluation exercise. Our proposed system was designed for addressing the problem of Author Profiling that consists of two subtasks, namely *author categorization* and *author ranking*. The former deals with detecting author's activity domain, e.g. journalist, authority, activist, company, celebrity, etc. The later consists in discovering which author have more influence (opinion maker) and which of those are less influential or have no influence at all (non opinion maker) among a community.

We proposed a supervised method for tackling the *author categorization* task. The main idea of such method was to perform attribute selection techniques based on document frequency scores and the transition point strategy. This method relies only in the information that authors provide on their Twitter profile. In its tuning phase we found that using only the 80% of the total number of attributes is enough for performing well

in this task. An interesting conclusion of the obtained results on the test set is that using term expansion by synonyms of the words appearing in the profiles is not useful. This might be because profile's text are usually not well written, thus the expansion only add noisy data to an already noisy description.

As future work for our proposed method for author categorization we plan to extend it by generating an activity domain model using external resources such as WordNet or Wikipedia. The idea of this modification is to generate enough (quality) information that can be used as a prototype description for each category or domain. Accordingly, we believe that terms expansion may perform better than our previous experiments since, in this case, terms to be expanded would be far less noisy than profile descriptions.

For the *author ranking* subtask we proposed a two-step chained method that consists of several processes. The first phase is a supervised approach, and the later a unsupervised approach that uses a Markov Random Field. A key element within our chain are the features extraction processes. Computed characteristics aim at capturing stylistic and behavioral features from authors. Our intuitive idea for proposing such features is that *opinion maker* users would have similar writing styles as well as similar posting behaviors. Some of the advantages of our proposed method is that it represents a fast and not expensive technique to determine an initial rank for authors. Once an initial rank has been provided, the MRF is able to improve the authors' ranking considering some stylistics and behavioral attributes.

Obtained results indicate, to some extent, that opinion makers do have similar writing styles as well as similar behavior patterns when using Twitter. In addition, obtained results also indicate that features extracted from profile information are insufficient and even noisy for the MRF configuration, however these represent a cheap and efficient option for assigning an initial ranking for authors. As future directions we plan to define and include an *observation potential* as proposed by [12]. By means of such potential, some information that is associated to a single node/user could be incorporated to the MRF and consequently a better distinction of opinion makers could be performed.

References

1. Enrique Amigó, Jorge Carrillo-de-Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Edgar Meij, Maarten de Rijke, and Damiano Spina. Overview of replab 2014: author profiling and reputation dimensions for online reputation management. In *Proceedings of the Fifth International Conference of the CLEF Initiative*, September 2014.
2. Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. A general evaluation measure for document organization tasks. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 643–652, New York, NY, USA, 2013. ACM.
3. Daria Bogdanova, Paolo Rosso, and Thamar Solorio. Modelling fixated discourse in chats with cyberpedophiles. In *EACL 2012 Workshop on Computational Approaches to Deception Detection*, pages 86–90, Avignon, France, 2012. ACL.
4. Ferro N.-Halvey M. Cappellato, L. and editors (2014) Kraaij, W. Clef 2014 labs and workshops, notebook papers. In *CEUR Workshop Proceedings (CEUR-WS.org) ISSN 1613-0073*, volume 1180, 2014.

5. Shoubin Kong and Ling Feng. A tweet-centric approach for topic-specific author ranking in micro-blog. In *Proceedings of the 7th International Conference on Advanced Data Mining and Applications - Volume Part I*, ADMA'11, pages 138–151, Berlin, Heidelberg, 2011. Springer-Verlag.
6. Gilberto Leon-Martagón, Esaú Villatoro-Tello, Héctor Jiménez-Salazar, and Christian Sánchez-Sánchez. Análisis de polaridad en twitter. *Journal of Research in Computing Science*, 62:69–78, 2013.
7. Dong Liu, Quanyuan Wu, and Weihong Han. Measuring micro-blogging user influence based on user-tweet interaction model. In Ying Tan, Yuhui Shi, and Hongwei Mo, editors, *Advances in Swarm Intelligence*, volume 7929 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013.
8. A. Pastor López-Monroy, Manuel Montes y Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, and Esaú Villatoro-Tello. INAOE's participation at PAN'13: Author proling task. In *Notebook for PAN at CLEF 2013*. Valencia, España, 2013.
9. M. F. Porter. Readings in information retrieval. chapter An Algorithm for Suffix Stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.
10. Berenice Reyes-Aguirre, Edgar Moyotl-Hernández, and Héctor Jiménez-Salazar. Reducción de términos índice usando el punto de transición. *Research on computing science*, 3:127–130, 2003.
11. Upendra Sapkota, Thamar Solorio, Manuel Montes y Gómez, and Gabriela Ramírez-De-La-Rosa. Author profiling for english and spanish text. In *Notebook for PAN at CLEF 2013*. Valencia, España, 2013.
12. Esaú Villatoro, Antonio Juárez, Manuel Montes, Luis Villaseñor, and L. enrique Sucar. Document ranking refinement using a markov random field model. *Natural Language Engineering*, 18(2):155–185, March 2012.
13. G.K. Zipf. *Human behavior and the principle of least effort: an introduction to human ecology*. Addison-Wesley Press, 1949.